DOCUMENT RESUME

ED 380 476 TM 022 780

AUTHOR Wright, Benjamin D.
TITLE Rasch Factor Analysis.
PUB DATE Oct 94
NOTE 34p.; Expanded version of a paper presented at the Annual Meeting of the Midwestern Educational Research Association (October 12-15, 1994).
PUB TYPE Reports - Evaluative/Feasibility (142) -- Speeches/Conference Papers (150)

EDRS PRICE MF01/PC02 Plus Postage.
DESCRIPTORS Comparative Analysis; *Estimation (Mathematics); *Factor Analysis; Instructional Leadership; *Item Response Theory; Principals; Problem Solving; *Rating Scales; Statistical Analysis; Teachers; Test Interpretation
IDENTIFIERS *Linear Models; *Rasch Model

ABSTRACT

Factor analysis and Rasch measurement are compared, showing how they address the same data with different interpretations of numerical status. Both methods use the same estimation method, with different measurement models, and they solve the same problem, with different utility. Factor analysis is faulted for mistaking stochastic observations of ordered labels as established linear measures and for failing to construct linear measurement. Using the Rasch measurement to replace factor analysis is developed for a dichotomy and shown for a rating scale example using responses of 2,049 Chicago (Illinois) public school teachers on the 13-item "Strength of Principal Leadership Scale" rating scale. Includes 11 figures. (Contains 18 references.) (Author/SLD)

RASCH FACTOR ANALYSIS[1]

Benjamin D. Wright
MESA, Education, University of Chicago

Abstract

Compares factor analysis and Rasch measurement. Shows how they: 1) address the same data, with different interpretations of numerical status, 2) use the same estimation method, with different measurement models, 3) solve the same problem, with different utility. Factor analysis is faulted for 1) mistaking stochastic observations of ordered labels as established linear measures and for 2) failing to construct linear measurement. How to use Rasch measurement to replace factor analysis is developed for a dichotomy and shown for a rating scale example.

Factor Analysis

Input datum x_{ni} is a test score, Likert rating or MCQ response of persons n=1,N to items (or tests) i=1,L. The raw data are expected to be sufficiently linear to allow equating incommensurable item origins and scales by subtracting local means and dividing by local standard deviations.

The sample standardized data for Factor 1 become:

$$z_{ni1} = (x_{ni} - m_i)/s_i \tag{1}$$

with $$m_i = \sum_n^N x_{ni}/N \qquad s_i^2 = \sum_n^N (x_{ni} - m_i)^2/(N-1)$$

This item scale equating expects complete data. Only persons with usable responses to every item can participate. When data are missing, they must be feigned or incomplete persons (or items) deleted. Deleting persons alters the interpretation of the standardizing sample. Deleting items alters the construct. Pair-wise deletion to estimate correlations biases factors.

The model for Factor 1 is:

$$z_{ni1} = u_{n1}v_{i1} + e_{ni1} \qquad e_{ni1} \sim N(0,\sigma_1^2) \tag{2}$$

$\{u_{n1}$ n=1,N$\}$ is a vector of person "factor scores", predicted for persons by Factor 1.
$\{v_{i1}$ i=1,L$\}$ is a vector of item "factor loadings", the regressions of $\{u_{n1}\}$ on the data from items i=1,L.

[1]A short version of this mss. was delivered at the Mid-Western Educational Research Association Annual Meeting, October 13, 1994.

The residual from Factor 1 is:

$$z_{ni1} - u_{n1}v_{i1} - e_{ni1} \rightarrow z_{ni2} \tag{3}$$

Whether this residual is all error $e_{ni1} \sim N(0,\sigma^2)$

or in addition to error implies other factors is presumably unknown.[2] When an additional factor is suspected, Factor 1 residuals are used for seeking Factor 2 and so on. Matrices of decreasing residuals $\{\{z_{nij}\}\}$ for j=2,M are extracted in turn to calculate the M factor model (Thurstone, 1947):

$$z_{ni1} - \sum_{j=1}^{M} u_{nj}v_{ij} + e \qquad e \sim N(0,\sigma^2) \tag{4}$$

Since there is no objective basis for a "right" number of factors nor a "theoretical" value for "error" σ^2 to factor down to, factor analysts default to conventions like: Stop when factor size (sum of squared factor loadings) becomes less than one. Stop when successive factor sizes level off. Stop after two or three factors, because anything more complicated is impossible to replicate.

The simplest way to obtain optimal values for person and item vectors $\{u_{nj}\}$ and $\{v_{ij}\}$ for each Factor j is to minimize:

$$\sum_{n=1}^{N}\sum_{i=1}^{L}(z_{nij} - \hat{u}_{nj}\hat{v}_{ij})^2 \tag{5}$$

This "direct factor analysis" (Saunders 1950, Cattell 1952, MacRae 1960, Wright and Evitts 1961) is a principal component decomposition (Hotelling 1933) of a "sample standardized" data matrix into j=1,M item vectors $\{v_{ij}\ i=1,L\}$ and M corresponding person vectors $\{u_{nj}\ n=1,N\}$.[3] The results are comparable to Thurstone centroids (1947).

Decomposition to identify each Factor j = 1,M is accomplished by initializing at $u_{nj}=1$ for all n and iterating through equations:

$$v_{ij} - \sum_{n=1}^{N} u_{nj}z_{nij}/N \qquad u_{nj} - \sum_{i=1}^{L} v_{ij}z_{nij}/\sum_{i=1}^{L} v_{ij}^2 \tag{6}$$

renormed by $C^2 - \sum_{n=1}^{N} u_{nj}^2/N \qquad u_{nj} - u_{nj}/C \qquad \text{so that} \qquad \sum_{n=1}^{N} u_{nj}^2 - N$

until successively smaller changes in $\{u_{nj}\}$ become uninteresting.

---

[2] Should another factor be expected, the most efficient approach is to analyze each subset of expected-to-be-singular items separately and to defer comparing any resulting "variables" until their construct definition and quantitative representation is established.

[3] Principle component decomposition is the core of most contemporary factor analysis and multi-dimensional scaling programs.

Standardized factor score u_{nj} is the value predicted by Factor j's regression on "independent" variables $i = 1,L$ with regression coefficients $\{v_{ij}\}$.

Factor j results are:

a) Factor j Size (variance "explained"):

$$G_j = \sum_{i=1}^{L} v_{ij}^2 \qquad \text{with} \qquad M < \sum_{j=1}^{M} G_j < L \tag{7}$$

b) Factor j Loadings (regression coefficients):

$v_{ij} = \sum_{n=1}^{N} u_{nj} z_{nij} / N$ of factor scores u_{nj} on residuals $\{\{z_{nij}\}\}$.

c) Factor j Standard Scores (zero mean, unit variance):

$u_{nj} = \sum_{i=1}^{L} v_{ij} z_{nij} / G_j$ predicted by regression coefficients $\{v_{ij}\}$

from residuals $\{\{z_{nij}\}\}$.

Problems with Factor Analysis

1. Raw data $\{\{x_{ni}\}\}$ are never linear measures. Even test scores, unless transformed into logits, become increasingly non-linear as they near their finite limits. When x_{ni} is a Likert rating it is not even cardinal! But ordinal data are not suited to Equations (1) through (5) and the factor scores they produce are necessarily non-linear.

2. In view of the non-linearity, the "true score" error models of Equations (2) and (3) do deal with uncertainty in a useful way.

3. The necessity for complete data is awkward. Data are never complete.

4. After each factor is extracted, its residuals $\{\{z_{nij}\}\}$ are the data for smaller factors. These residuals contain one sure effect, the turbulence left behind by the estimation of preceding factors. Intimations of smaller factors are necessarily awash in the residual wake of larger factors.

5. Without a basis for anticipating a final "error" size for σ^2, there is no objective way to decide when to stop factoring.

6. Software implementations seldom provide standard errors for factor loadings or factor scores.

7. When a "same" set of items is refactored from a new sample of persons, neither factor sizes nor loadings are ever the same. Only the most generous fudging allows one to suppose their factor structure has been confirmed.

Most factor analysts swallow problems 1 through 6. But problem 7 is fatal. As the numerical instability of presumed "replications" emerges, we are forced to retreat from non-reproducible numbers to nominal conclusions. Person factor scores are abandoned. All but the relative magnitudes of item factor loadings is ignored. The only use we make of the factor analysis output is to classify items according to their highest factor loading. Person scores for each category of items are obtained, not from factor scores (even when separate refactorings are done for each class of items) but by summing the original standardized (but inevitably non-linear) person responses z_{ni1} to the items in a factor class. Whatever the distribution of item loadings in a factor class may suggest, all items are given equal weight in this summation of non-linear numerical labels.

Rasch Factor Analysis

Why not admit that our data are neither measures nor cardinal numbers but necessarily begin as nothing more than labels $\{\{c_{ni}\}\}$ for nominal qualities - labels which may respond to an intelligent ordinal scoring x_{ni}= 0,1,2,3,, to produce the ordinal score matrix $\{\{x_{ni}\}\}$?

Familiar examples are: **rating scales** like (*strongly agree > agree > disagree > strongly disagree*) which can usually be scored x_{ni} = 3,2,1,0 or at least x_{ni} = 2,1,0,0 and **MCQ options** like (*right > wrong*) which can always be scored x_{ni} = 1,0. Even **raw scores** ($r+1 > r > r-1$) can respond to an ordinal interpretation.

Why not embrace the inescapable **initially nominal but possibly ordinally interpretable** status of datum x_{ni} and address it directly for what it is with a probability model for the occurence of ordered categories (Rasch 1961, Andrich 1978, Wright and Masters 1982)?

First, we will show the algebra of this approach in its simplest form, the use of x_{ni}=0,1 to represent a dichotomy through which nominal events interpreted as signifying "more" of an intended variable are scored "1" and nominal events signifying less are scored "0" (Rasch 1960/1980/1992, pp 62-124).

Then we will illustrate the empirical similarities and differences between factor analysis and Rasch measurement by analyzing the responses of 2049 public school teachers to a "Strength of Principal Leadership" rating scale.

To begin the algebra for $x_{ni}=0,1$ we will not mistake the number labeling as a linear measure, but will, instead, recognize it for the binomial it is - a stochastic binomial for which we have decided, on the basis of our measuring intentions, what kind of events are "better" for our purposes than others, i.e. what we decide qualifies as a "right" answer. To set up a counting system for this preference, we label the preferred ("right" answer) event "1" and it's absence ("not right" hence "wrong" answer) "0".

The error model which follows is not the ill-suited linear error (true score) model of factor analysis (Equation 2) but a binomial probability model P_{nix} for the occurrence of $x_{ni}=0,1$ (Equation 8) which, because of its formulation (Rasch 1960):

1) Obtains the parameter separability necessary for constructing objective (additive conjoint) measurement (Luce and Tukey 1964, Perline, Wright and Wainer 1979) and

2) Has Fisher sufficient statistics (1922) for linear person measures B_n and item calibrations D_i which combine additively as (B_n-D_i) (and therefore construct the linearity we need for subsequent quantitative analysis) to govern the probabilities of $x_{ni} = 0,1$

The necessary and sufficient model is:

$$\log(P_{ni1}/P_{ni0}) = B_n-D_i \qquad x_{ni} = 0,1 \tag{8}$$

from which

$$E_{xni} = P_{ni1} \qquad V_{xni} = P_{ni1}(1-P_{ni1}) = P_{ni1}P_{ni0}$$

Parameters for this model, as with factor analysis (5), are estimated by minimizing:

$$\sum_{n=1}^{N}\sum_{i=1}^{L}(x_{ni}-\hat{E}_{xni})^2$$

but now $$\hat{E}_{xni} = \hat{P}_{ni1} = \exp(\hat{B}_n-\hat{D}_i)/[1+\exp(\hat{B}_n-\hat{D}_i)] \tag{9}$$

rather than $$\hat{E}_{zni} = \hat{U}_n\hat{V}_i$$

With this Rasch approach to x_{ni} we get not only least-square (the implementation here for maximizing likelihood) estimates of linear person measures B_n and linear item calibrations D_i along the single common measurement line (variable) which they conjointly define, but also a stochastic basis, the binomial error variances $p_{ni1}p_{ni0}$, for estimating relevant standard errors for B_n and D_i and for evaluating the probabilities of residuals:

$$y_{ni} = x_{ni} - p_{ni1} \qquad z_{ni} = y_{ni}/\sqrt{p_{ni1}p_{ni0}} \qquad E_{zni} = 0 \qquad V_{zni} = 1 \tag{10}$$

This enables a detailed misfit analysis which, in turn, allows a partition of the full matrix of residuals $\{\{z_{ni1}\}\}$ into those many z_{ni1} which are observed to be no greater than the probability model expects, say $|z_{ni1}| < 2$, and hence, in "all probability", of no immediate empirical interest and the, possibly interesting, subset of remaining, more extreme residuals, say $|z_{ni1}| > 3$, which are sufficiently improbable to invite further investigation and reconsideration as possible evidence of a second variable.

It is only when improbable residuals emerge that there is an empirical incentive to look for a presumably unexpected second variable.[4]

Should we decide to venture further, the improbable residuals tell us exactly where to look. No need to accumulate confusion by wallowing through the full matrix $\{\{z_{ni1}\}\}$ or by subsequent factor rotations which are, after all, directed by the largest residuals. We already know from the residuals in hand which particular items (and also which persons) do not fit into the construction of the first variable. Should there be another useful, albeit unexpected, variable in these data, it will be most directly accessible among the original (rather than residual) person responses to the subset of items which misfit the first analysis.

To seek a second variable j=2, therefore, we concentrate on just the data for the subset of misfitting items.[5] We apply the Rasch probability model again, not to the whole matrix of residuals $\{\{z_{ni1}\}\}$, but only to the submatrix $\{\{x_{ni}, i \in j=2\}\}$ of original ordinally interpreted responses of persons to this subset of items. We estimate a new set of linear item calibrations $\{D_{i2}\}$ for just these misfitting items along with a new set of additive conjoint person measures $\{B_{n2}\}$ on this newly defined "Variable 2".

---

[4] In a sensible research, of course, a "theoretical" incentive would dominate most "empirical" results. There would be a set of well-designed items which were intended to define a single variable. The "empirical" question would narrow to finding out whether any of these well-intended items failed to perform usefully and, if so, with which particular persons and possibly "Why?".

[5] The focusing provided by identifying items manifesting improbable residuals is analogous in purpose and consequence to the item clustering by which factor rotation is guided.

To find out whether unexpected Variable 2 brings out any new information, we plot Variable 2 person measures $\{B_{n2}\}$ against Variable 1 person measures $\{B_{n1}\}$. The shape of this plot shows us in detail the extent to which we have found a useful second variable and also for which of these persons it may actually provide new information.

Because we have independent standard errors for each linear measure on each variable, the statistical status of the differences $(B_{n2} - B_{n1})$ for each of the n=1,N persons can be judged objectively by comparing them with their estimated standard errors:

$$(B_{n2} - B_{n1}) / \sqrt{SE_{n2}^2 + SE_{n1}^2} = T_{n21} \sim N(0,1) \tag{11}$$

Extension of the dichotomous Rasch model to the ordered response categories $x_{ni}=0,1,2,3,,,m,,,\infty$ of rating scales, partial credits, grades, ranks, raw scores, counts and to models with additional facets for raters and tasks is straightforward (Wright & Masters 1982, Linacre 1989).

An Empirical Example

We will illustrate the empirical similarities and differences between factor analysis and Rasch measurement by using both methods to analyze the responses which 2049 Chicago public school teachers gave to the 13 item "Strength of Principal Leadership" rating scale on page 8 of the *Consortium on Chicago School Research* questionnaire "Charting Reform: The Teachers' Turn, 1994".[6]

The 13 rating items in Figure 1 were written to define a single line of inquiry which produced a single measure of perceived "Strength of Principal Leadership" for each of the 2049 teachers. The methodological question then is: Are these 13 items used by these 2049 teachers in a way that enables the construction of a reasonable and useful single measure?

[Figure 1]

The three items about to be exposed as diverging from the coherent core defined by the other 10 are marked [A], [B] and [C] in Figure 1.

Figure 2 is the **factor analysis scree plot** of principal component eigenvalues for these data. Some factor analysts might conclude, at this point, that the 13 items work together well enough to define a single 13 item factor and stop. The scree plot, however, does hint that components [2] and [3] may be a bit too large.

---

[6]The research from which these data come is supported by the *Consortium on Chicago School Research* under the direction of Anthony S. Bryk and Penny Sebring. The factor analyses were done with SAS by Stuart Luppescu. The Rasch measurement analyses were done with BIGSTEPS by Winifred Lopez.

[Figure 2]

Figure 3 is a **Rasch measurement item misfit plot** for the same data. Here the salient misfit of items [A], [B] and [C] is unequivocally distinct.

[Figure 3]

Figure 4 is the table of **varimax factor loadings**. Factor analysts who rotate these data will not miss the exceptionality of items [A], [B] and [C] and, after studying their item text, will have some useful ideas as to why these three items might not follow the mainline defined by the 10 item core.

[Figure 4]

Figure 5 is the **Rasch measurement item calibration table** listed both in misfit and in measurement orders. Here we see for each of the 13 items not only its fit statistics in mean square and standardized form but also its raw score point-biserial and relative "difficulty to be agreed with", its "unpopularity" as it were.

[Figure 5]

Figure 5 shows that items [B] and [C] are 0.4 logits harder to agree with (1.04 and 0.99 versus 0.61 logits) than the next "hardest" item. Their texts share a close, personal supervision of teacher by principal. This supervision could be supportive but it is more likely to be restrictive. Indeed the verb "supervises" is viewed by many as counter-collaborative. Thus these items are not only hard to agree with but also ambiguous with respect to the spirit of increasing egalitarian collaboration which dominates the hierarchy of the 10 core items.

Item [A], "Principal makes all final decisions." at the other end of the line (at -0.77 logits) is the item easiest to agree with but also emphatically counter-collaborative.

Once the text of these three items is examined and understood in terms of the pro-collaborative tenor of the 10 core items, it is easy to agree with both factor analysis and Rasch measurement results and to remove these three items from the definition of this "Strength of Principal Leadership" variable.

Indeed, at this point we may find ourselves wondering why we included these three items in the first place. We may also wonder, now that we see the construct evolution of the variable defined by the 10 core items, whether it would not be useful to rename this variable "Strength of Principal Support for Collaboration".

Do you see how useful the Rasch measurement information in Figure 5 can be for confirming a decision to set aside the three aberrant items and to identifying the construct evolved by the hierarchy of the 10 core items? Other parts of the standard Rasch measurement output are equally useful.

Figure 6 maps both 2049 teachers and 10 core items onto the single line of inquiry that the 10 items define. The line of inquiry rises from low agreement at the bottom of Figure 6 up to high agreement at the top. On the left, each teacher is located at their level of agreement. On the right, each item is located first at its level of transition from "strongly disagree" to "disagree", then at its level of transition from "disagree" to "agree" and finally, on the far right, at its level of transition from "agree" to "strongly agree".

[Figure 6]

Figure 7 shows the same data in a different form. Now the line of inquiry is drawn to increase from left to right. The exact count of teachers at each level of agreement is given along the bottom of the upper figure. We can see that 127 teachers at the top of Figure 6 and at the far right of Figure 7 have "strongly agreed" with all 10 items. We can also see that the modal group of 333 teachers at a measure of about 1.3 logits is above the disagree-to-agree transition of even the hardest to agree with item. And we can see that 59 teachers at the far left of Figure 7 have claimed to "strongly disagree" with all 10 items.

[Figure 7]

Figure 7 also shows us something of considerable importance to our understanding and application of this rating scale as it was used by these teachers. The spacing beween rating scale categories (1) to mark "strongly disagree, (2) to mark "disagree", (3) to mark "agree" and (4) to mark "strongly agree" is not uniform (equal) as a Likert interpretation would have it. Values for the expected difficulties of each step are given in the table at the bottom of Figure 7. The estimated increase in difficulty from "strongly disagree" at step (1) and "disagree" at step (2) in expected step measures is $-1.74 - (-3.74) = 2.0$ logits. But the estimated increase from "agree at step (3) and "strongly agree" at step (4) is $4.57 - 1.28 = 3.29$ logits. The second distance is 1.65 times greater than the first. We see that it is tangibly easier to move up from "strongly disagree" to "disagree" and so reduce strong disagreement, than it is to move up from "agree to "strongly agree", and so produce strong agreement. Figure 7 also shows us why we might prefer to avoid factoring Likert scores like 1,2,3,4 as though they were equally spaced measures.

Rasch measurement also shows us useful information about each of the 2049 teachers. Figure 8 begins this part of the Rasch analysis by showing the distribution of teacher response pattern misfit. We see that a substantial number of teachers are using these 10 items idiosyncratically. Subsequent pages of output identify these teachers and show on which items they provided surprising ratings. These diagnostic outputs show us how teachers use the 10 items and differentiate the many teachers whose responses are sufficiently coherent to produce a valid measure of perceived "Strength of Principal Support for Collaboration" from other teachers whose response patterns make unique individual statements.

[Figure 8]

Figure 9 shows the exact and non-linear relationship between factor scores and Rasch measures for these 2049 teachers. Since the Rasch measures are modelled to be linear and the Rasch fit analyses expose any failures of data to support the linear measure construction based on this modelling, it must be the factor scores which are not linear.

[Figure 9]

Figure 10 summarizes Richard Smith's use of two-factor simulated data to evaluate how well factor analysis and Rasch measurement detect a second factor. Smith finds that factors of equal size can only be discerned when they are uncorrelated $R<.3$. Against that kind of data factor analysis does better than Rasch measurement.

[Figure 10]

Against all other kinds of data, however, particular the kind of data most frequently encountered in social science research in which the factors are NOT of equal size and NOT uncorrelated, i.e. the usual situation where there is first an intended dominant factor and then an unintended off-shoot, correlated with the first factor and less well represented, Rasch measurement does better.

Finally, Figure 11 collects into one summary table the considerations which bring out the similarities and differences between factor analysis and Rasch measurement.

[Figure 11]

References

Andrich, D. (1978). A rating formulation for ordered response categories. Psychometrika, 43, 561-573.

Cattell, R.B. (1952). Factor Analysis. New York: Harper, 414-416.

Fisher, R.A. (1922). On the mathematical foundations of theoretical statistics. Philosophical Transactions of the Royal Society of London (A), 222: 309-368.

Hotelling, H. (1933). Analysis of a complex of statistical variables into principal components. Journal of Educational Psychology, 24, 417-441, 498-520.

Linacre, J.M. (1989). Many-Facet Rasch measurement. Chicago: MESA Press.

Luce, R.D. & Tukey, J.W. (1964). Simultaneous conjoint measurement. Journal of Mathematical Psychology, 1, 1-27.

Lumsden, J. (1961). The construction of unidimensional tests. Psychological Bulletin, 58, 122-131.

MacRae, D. (1960). Direct factor analysis of sociometric data. Sociometry, 23, 360-371.

Perline, R.,Wright, B.D. and Wainer, H. (1979). The Rasch model as additive conjoint measurement. Applied Psychological Measurement, 3, 237-256.

Rasch, G. (1960/1980/1993). Probabilistic Models for some Intelligence and Attainment Tests. Chicago: MESA Press.

Rasch, G. (1961). On general laws and the meaning of measurement in psychology. Proceedings of the Fourth Berkeley Symposium on Mathematical Statistics and Probability, 4, 321-33. Berkeley: University of California Press.

Saunders, D.R. (1950). Practical Methods in the Direct Factor analysis of Psychological Score Matrices. Ph.D. Dissertation. Urbana: University of Illinois.

Smith, R.M. and Miao, C.Y. (1991). Assessing unidimensionality for Rasch measurement. In M. Wilson (ed.) Objective Measurement: Theory into Practice, Vol. 2. Norwood, N.J.: Ablex Publishing Corp.

Thurstone, L.L. (1947). Multiple-Factor Analysis. Chicago: University of Chicago Press.

Wright, B.D. and Linacre, J.M. (1989/94). BIGSTEPS: Rasch Analysis Computer Program. Chicago: MESA Press.

Wright,B.D. and Evitts, M. (1961). Direct factor analysis in sociometry. Sociometry, 24, 82-98.

Wright, B.D. and Masters, G.N. (1982) Rating Scale Analysis. Chicago: MESA Press.

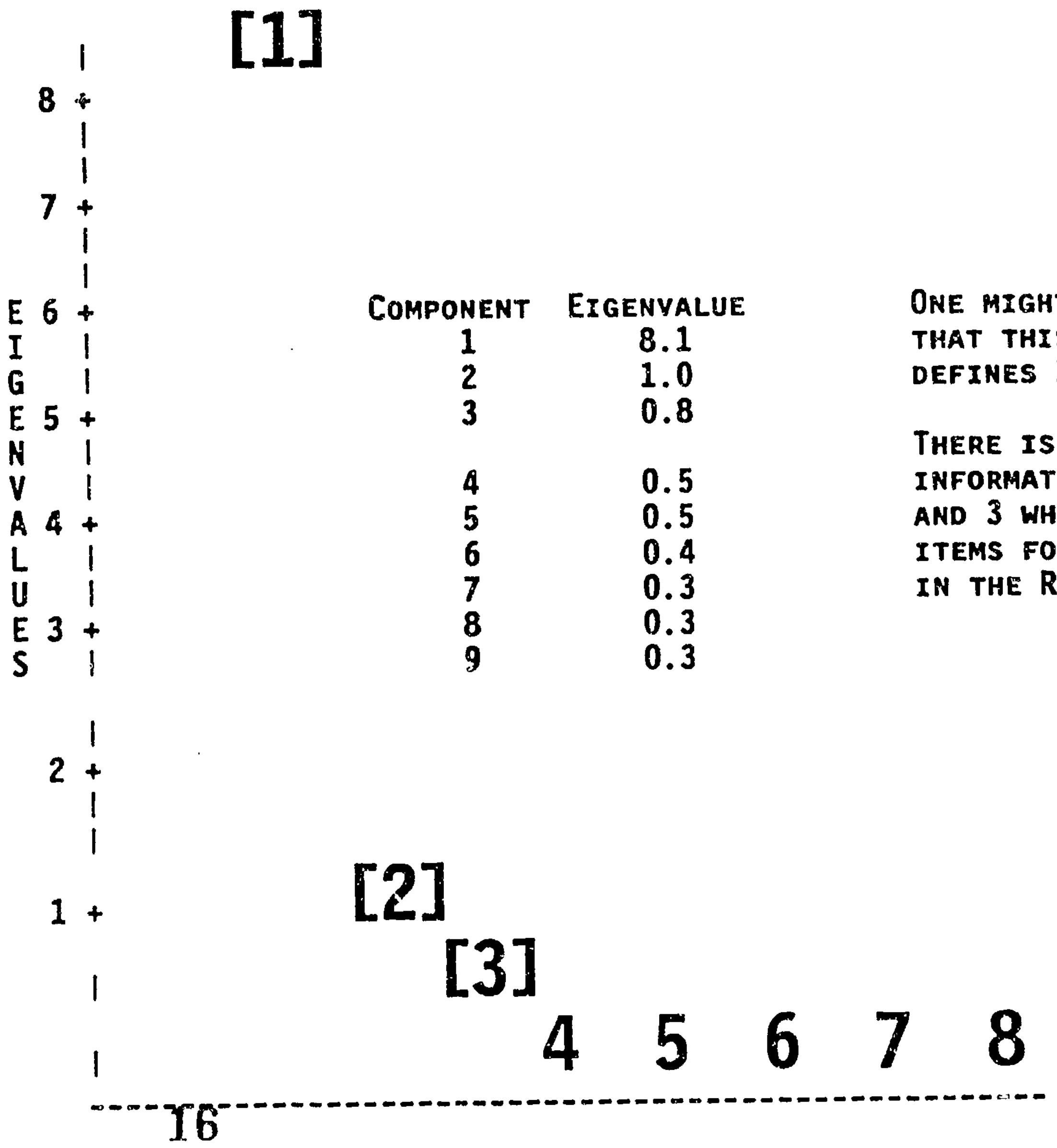
Wright, B.D. and Stone, M. (1979). Best Test Design. Chicago: MESA Press.

Figure 1
The Intended Variable is:
STRENGTH OF PRINCIPAL LEADERSHIP
To be constructed from teachers' responses
to the following 13 items.

The principal at this school: strongly		strongly disagree	disagree	agree	agree
Encourages teachers to try new methods of instruction		()	()	()	()
Is willing to make changes		()	()	()	()
Makes clear to the staff his or her expectations for meeting instructional goals		()	()	()	()
Sets high standards for teaching		()	()	()	()
Promotes parental and community involvement in the school		()	()	()	()
Supports and encourages teachers to take risks		()	()	()	()
Understands how children learn		()	()	()	()
Works to create a sense of community in this school		()	()	()	()
Communicates a clear vision for our school		()	()	()	()
Visits classrooms regularly	[B]	()	()	()	()
Makes the final decision on all important matters	[A]	()	()	()	()
Is strongly committed to shared decision making		()	()	()	()
Closely supervises teachers' work	[C]	()	()	()	()

Consortium on Chicago School Research Questionnaire
"Charting Reform: The Teachers' Turn, 1994" Page 8

Figure 2
How Many Factors Are There?
SCREE PLOT OF EIGENVALUES
for 13 Original Items

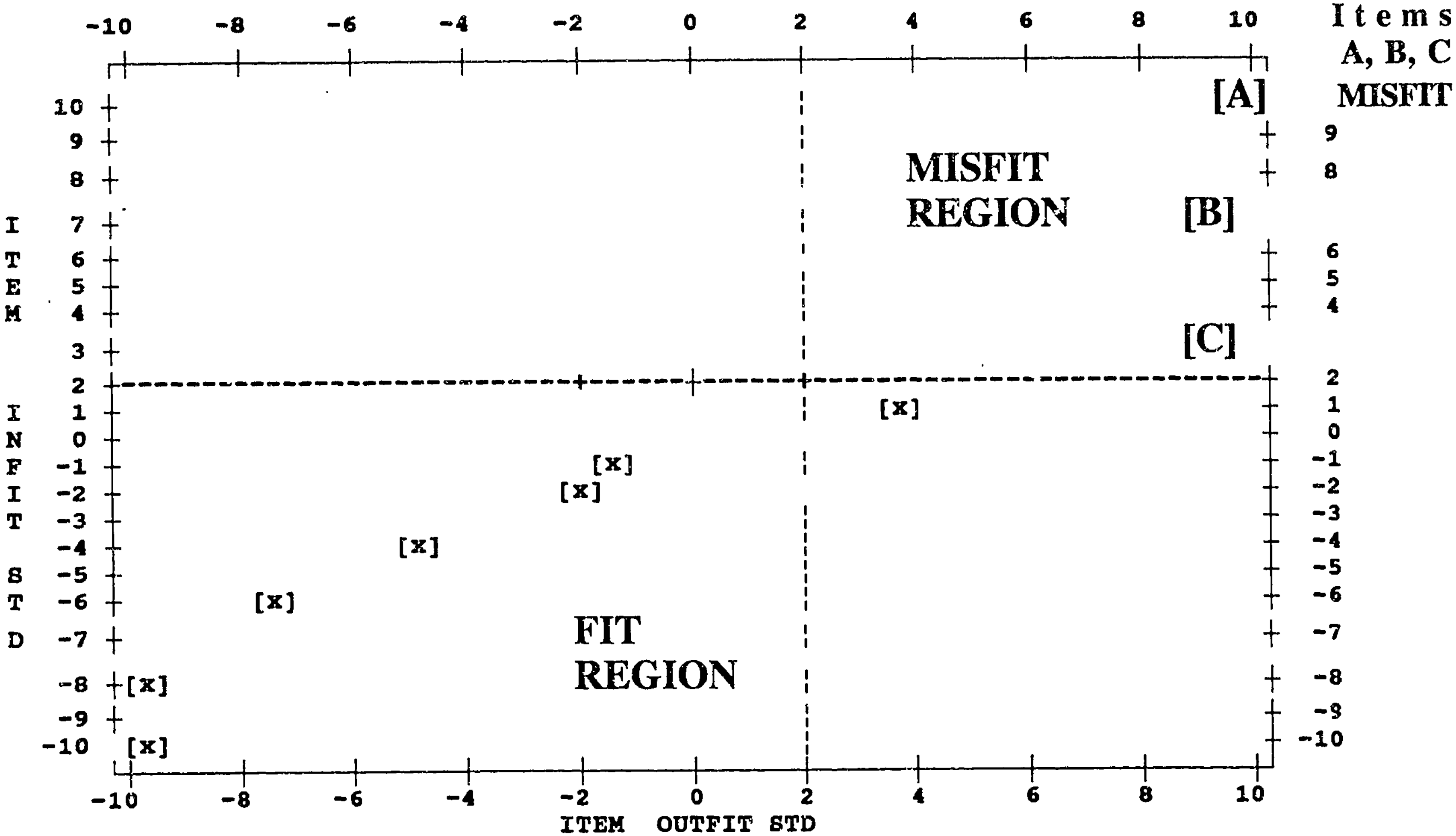


COMPONENT	EIGENVALUE
1	8.1
2	1.0
3	0.8
4	0.5
5	0.5
6	0.4
7	0.3
8	0.3
9	0.3

ONE MIGHT CONCLUDE THAT THIS SET OF 13 ITEMS DEFINES 1 AND ONLY 1 FACTOR.

THERE IS, HOWEVER, USEFUL INFORMATION IN ROTATED FACTORS 2 AND 3 WHICH ISOLATE THE SAME 3 ITEMS FOUND TO BE WORST FITTING IN THE RASCH ANALYSIS.

Figure 3
Which Items Misfit?
RASCH ITEM FIT PLOT
for 13 Original Items



Three items marked [A], [B] and [C] and identified by Rasch statistics and item text in Figure 5 are found to misfit more extremely than the remaining 10 items. These are the same 3 items identified as not on the main factor by the varimax factor rotation reported in Figure 4.

Figure 4

Which Items Are Not On The Main Factor?

ROTATED FACTOR LOADINGS for 13 Original Items

FACTOR F1	LOADINGS F2	F3	THE PRINCIPAL AT THIS SCHOOL:	Rasch MISFIT items:
0.81	0.15	0.12	encourages teachers to try new methods	
0.83	0.20	-0.00	is willing to make changes	
0.72	0.38	0.24	makes teaching expectations clear	
0.74	0.38	0.24	sets high standards for teaching	
0.75	0.26	0.15	promotes community involvement	
0.78	0.23	0.03	encourages teachers to take risks	
0.75	0.35	0.16	understands how students learn	
0.78	0.36	0.07	works to create community	
0.75	0.41	0.18	communicates a clear vision	
0.33	[0.85]	0.07	visits classrooms regularly	[B]
0.12	0.14	[0.95]	makes all final decisions	[A]
0.70	0.47	-0.05	committed to shared decisions	
0.33	[0.84]	0.20	closely supervises teachers	[C]

Variance explained: F1 = 6.1, F2 = 2.6, F3 = 1.2

Rotated Factors 2 and 3 find items [A], [B] and [C] not on the main factor.

Figure 5
Which Items Misfit?
RASCH MISFIT ORDER for 13 Original Items

MEASURE	ERROR	MISFIT MNSQ	MISFIT STD	PTBIS	THE PRINCIPAL AT THIS SCHOOL:	
-.77	.04	2.78	9.9	A .21	makes all final decisions	Factor 3
.99	.04	1.34	9.4	B .61	visits classrooms regularly	Factor 2
1.04	.04	1.32	8.9	C .62	closely supervises teachers	Factor 2
.57	.04	1.13	3.8	.69	encourages teachers to take risks	

MEASURE	ERROR	MNSQ	STD	PTBIS	THE PRINCIPAL AT THIS SCHOOL:	
1.04	.04	1.32	8.9	C .62	closely supervises teachers	Factor 2
.99	.04	1.34	9.4	B .61	visits classrooms regularly	Factor 2
.61	.04	.95	-1.5	.74	strongly committed to shared decisions	
.57	.04	1.13	3.8	.69	encourages teachers to take risks	
.05	.04	.66	-9.9	.83	communicates a clear vision for school	
.02	.04	.76	-7.5	.80	works to create community in school	
-.20	.04	.68	-9.8	.78	understands how children learn	
-.30	.04	.69	-9.6	.79	makes teaching expectations clear	
-.32	.04	.83	-4.7	.73	is willing to make changes	
-.38	.04	.63	-9.9	.81	sets high standards for teaching	
-.62	.04	.92	-2.1	.69	encourages teachers to try new methods	
-.69	.04	.73	-7.6	.74	promotes community involvement	
-.77	.04	2.78	9.9	A .21	makes all final decisions	Factor 3

Rasch misfit analysis finds the same 3 deviant items found by factor analysis. Qualitatively i.e. finding deviant items, factor analysis and Rasch analysis reach the same conclusion. But there are substantial differences in the utility and inferential stability of quantitative results. Factor analysis gives sample dependent item regression coefficients and the sample standardized person scores predicted by this regression on the data. Rasch analysis gives test-free person measures and sample-free item calibrations positioned together on the common linear scale defined by the items.

Figure 6
Rasch MAP OF TEACHERS on the 10 BEST ITEMS

```
MEASURE                                                                        MEASURE
        ------TEACHERS-+-ITEMS  - LOW  -+-ITEMS  - MEAN-+-ITEMS  - HIGH
  6.0    .###########  +                +               +                        6.0
                 .###
        Teachers     .
  5.0   who      .###  +                +               +                        5.0
        STRONGLY     .
        AGREE    .###                                     XX   midway
  4.0            .###  +                +               +      between           4.0
                     .                                    XX    AGREE
                  .##                                     X      and
  3.0            .###  +                +               + XXX  STRONGLY          3.0
                  .##                                     XX    AGREE
                  .##
  2.0                . +                +               +                        2.0
                 .###
      .############
  1.0                . +                + XX   midway   +                        1.0
               .#####                          between
                .####                     XX   DISAGREE
   .0             .##  +                + X       and   +                         .0
                   ##                     XXX    AGREE
                .####                     XX
 -1.0             .##  +                +               +                       -1.0
                  .##
                  .##
 -2.0   Teachers   .#  + XX   midway    +               +                       -2.0
        who        .#         between
        STRONGLY     . XX     STRONGLY
 -3.0   DISAGREE     . + X    DISAGREE  +               +                       -3.0
                     . XXX      and
                     . XX     DISAGREE
 -4.0                . +                +               +                       -4.0
        -------PERSONS-+-ITEMS  - LOW  -+-ITEMS  - MEAN-+-ITEMS  - HIGH
 EACH '#' IN TEACHER COLUMN IS  28 TEACHERS; EACH '.' IS 1 TO  27 TEACHERS
```

Rasch output of teacher measures in the metric of the variable defined by item calibrations enables MAPPING teachers and items together into a single picture showing the spread and targeting of the sample of persons on the span of defining items. These MAPS show normative and criterion validity simultaneously. MAPS also s' ɔw for each person their expected level of agreement with each item.

Figure 7
RASCH RATING SCALE SPACING ANALYSIS
on the 10 Best Items

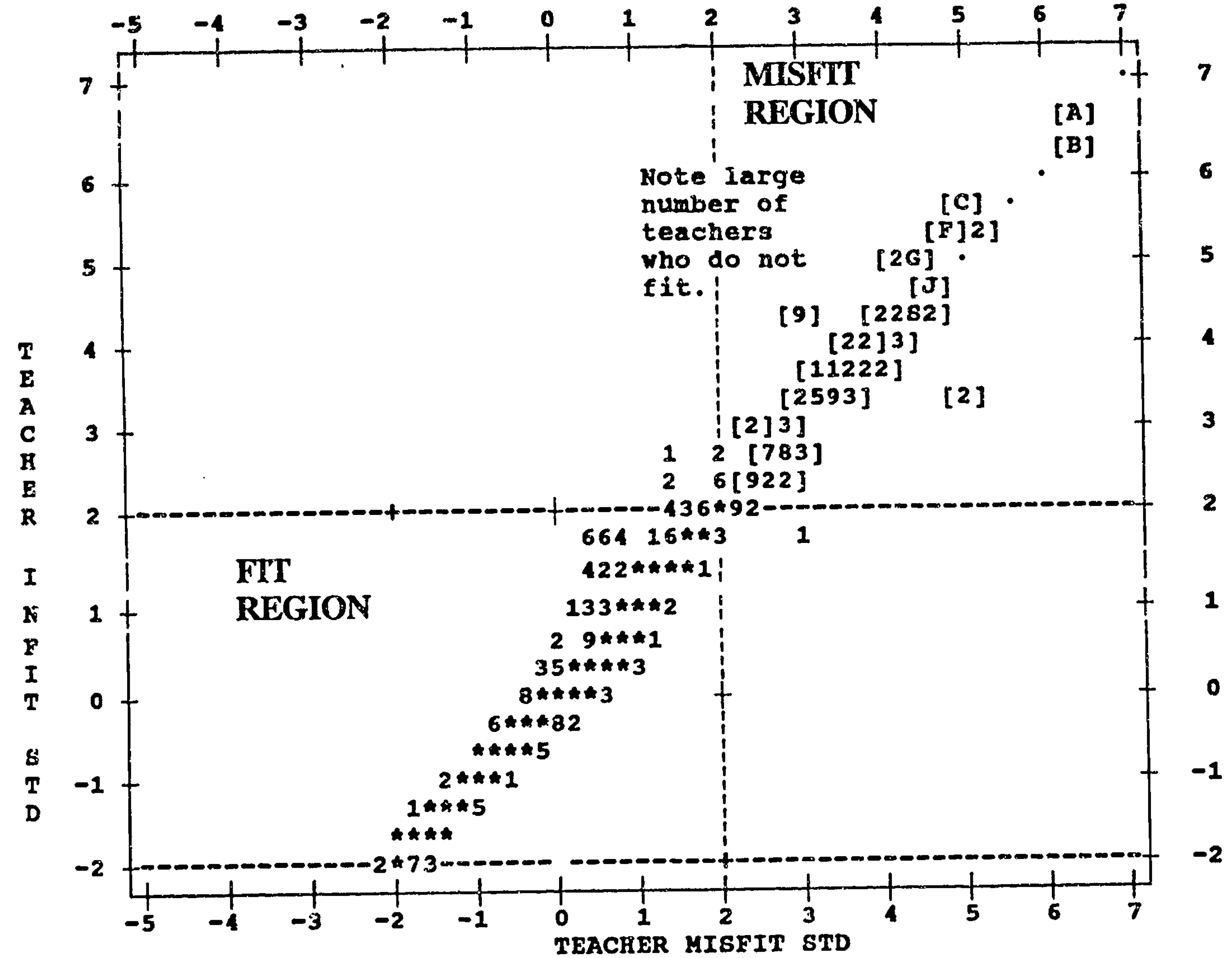
```
EXPECTED SCORE: MEAN   (":" INDICATES HALF-SCORE POINT)
-5        -3        -1         1         3         5         7  THE PRINCIPAL:
|---------+---------+---------+---------+---------+---------|
         . 1   :     2   :     .   3         :     4 .         is strongly committed
         . 1   :     2  :      .   3         :    4  .         supports encourages
       1 .  :    2     :       .3          :    4    .         communicates a vision
       1 .  :   2      :       3          :     4 .  .         works for community
      1  .:    2     :        3.         :    4      .         understands children
     1   .:    2     :       3 .        :     4      .         has clear expectations
     1   :     2     :       3 .        :     4      .         is willing to change
     1   :     2    :        3 .        :    4       .         has high standards
   1   : .   2     :      3    .       :    4        .         encourages new methods
   1   : .   2    :       3    .       :   4         .         promotes involvement
|---------+---------+---------+---------+---------+---------|
-5        -3        -1         1         3         5         7

                                1 1 3  1                     3
5   1   1 11 1122245 566147 115 32 0 7 6 815  8 9  9    8     2   TEACHER
99  4   4 00 989878283262458242333115729378239163176257 1     7   DISTRIBUTION
    Q            S            M           S           Q
```

RASCH SUMMARY OF RATING SCALE STEP MEASURES

CATEGORY LABEL	OBSERVED COUNT	STEP MEASURE	EXPECTED SCORE MEASURES STEP-.5	AT STEP	STEP+.5	THURSTONE THRESHOLD	AVG MEASURE
1	1314	NONE		[-3.74]	-2.87		-2.46
2	3219	-2.51	-2.87	[-1.74]	-.63	-2.67	-.74
3	10729	-.96	-.63	[1.28]	3.50	-.80	1.31
4	4713	3.46	3.50	[4.57]		3.47	3.98
		modal		mean		median	

The Likert ratings labeled 1,2,3,4 used by these teachers are NOT EQUALLY SPACED IN THEIR USE. The teachers needed only 2.00 logits to go from STRONGLY DISAGREE at -3.74 to DISAGREE at -1.74, but 3.29 logits to go from AGREE at 1.28 to STRONGLY AGREE at 4.57. Most Likert scales are even more unequally spaced in use. Instead of supposing equal spacing, Rasch analysis lets the responses of the persons using the rating scale determine the spacing actually in effect for them during their ratings.

Figure 8
FINDING MISFITTING TEACHERS
on the 10 Best Items



Each of these misfitting teachers can be recognized for what is unique in their views by examining the individual response vectors printed in BIGSTEPS Table 7 "Misfitting Persons". Many of the misfits appearing here are due to ethnic differences in the use of 2 of the 10 items. But that's another story.

Figure 9
FACTOR SCORES versus RASCH MEASURES
on the 10 Best Items for the 2049 Teachers

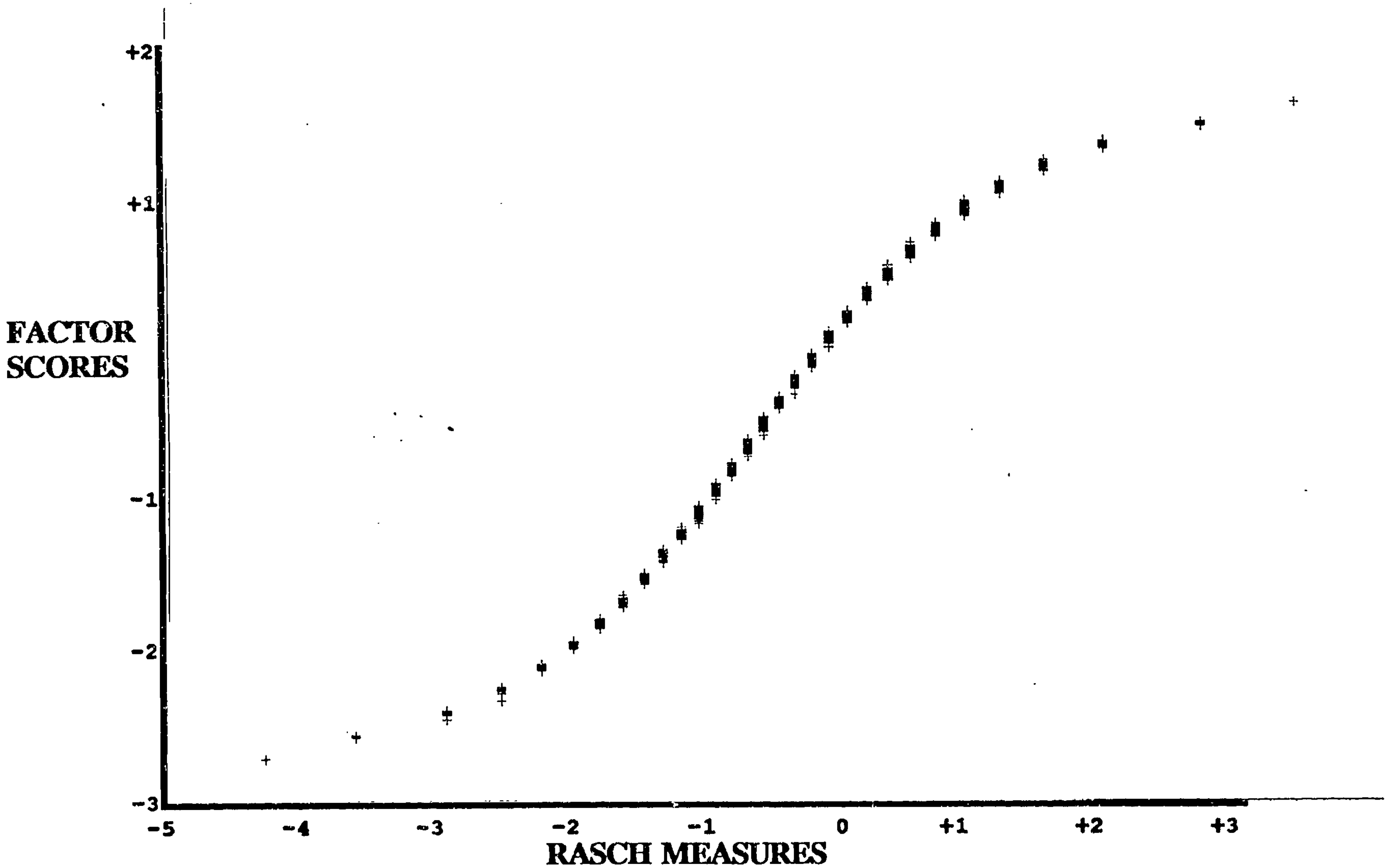


Figure 10

RASCH MEASUREMENT vs FACTOR ANALYSIS - in General

Smith, R.M. [(1992). Assessing unidimensionality for the Rasch Rating Scale Model. AERA, San Francisco, April 1992] evaluates the dimensionality sensitivity of principle component factor analysis (by SPSS-PC) and Rasch measurement (by BIGSTEPS) for 5-category rating scale data simulated from pairs of variously correlated factors x and y represented by equal and unequal numbers of items N_y and N_x. "Yes" for factor analysis means > 70% of the y items loaded on the 2nd factor. "Yes" for Rasch measurement means > 70% of the y items produced standardized mean square outfits > 2.0.

Rxy	Ny vs Nx		Factor Analysis		Rasch Item MISFIT
	5 - 35		Yes		Yes
	10 - 30		Yes		Yes
	15 - 25		Yas		Yes
R=.07	20 - 20	FA better	Yes		NO
	5 - 35		Yes		Yes
	10 - 30		Yes		Yes
	15 - 25		Yes		Yes
R=.26	20 - 20	FA better	Yes		NO
	5 - 35		Yes		Yes
	10 - 30		Yes		Yes
	15 - 25		Yes		Yes
R=.39	20 - 20		NO		NO
	5 - 35		Yes		Yes
	10 - 30		Yes		Yes
R=.55	15 - 25		NO	Rasch better	Yes
R=.55	20 - 20		NO		NO
	5 - 35		Yes		Yes
R=.72	10 - 30		NO	Rasch better	Yes
R=.72	15 - 25		NO	Rasch better	Yes
	20 - 20		NO		NO
R=.89	5 - 35		NO	Rasch better	Yes
R=.89	10 - 30		NO	Rasch better	Yes
	15 - 25		NO		NO
	20 - 20		NO		NO

Equal size factors
WHEN $N_y = N_x$
can only be separated
when nearly $R < .3$

UNCORRELATED

Rasch item MISFIT
does better than
Factor analysis at
separating factors
CORRELATED $R > .5$

WHATEVER $N_y < N_x$

Figure 11
COMPARING FACTOR ANALYSIS AND RASCH MEASUREMENT

ASPECT	FACTOR ANALYSIS	RASCH MEASUREMENT
MOTIVE	summarize data	construct measurement
INPUT MODEL	established linearities	stochastic events
MISSING DATA	loses rows or cols biases factors	routinely accommodated no data lost
EQUALIZATION	norm-standardized into $z_{ni} = (x_{ni}-m_i)/s_i$	criterion-classified into ordered categories $x_{ni} = 0,1,2,,,K_i$
METHOD	principal components	latent trait
MODEL	$z_{ni} = u_n v_i + e_{ni}$	$\log[P_{nix}/P_{nix-1}] = B_n - D_i - F_x$
ESTIMATED BY MINIMIZING	$\sum_{n=1}^{N}\sum_{i=1}^{L}(z_{ni}-\hat{u}_n\hat{v}_i)^2$	$\sum_{n=1}^{N}\sum_{i=1}^{L}(x_{ni}-\hat{E}_{xni})^2$
ANCHORED AT	$\sum_{n=1}^{N} u_n^2/N = 1$	$\sum_{i=1}^{L} D_i = 0 \quad \sum_{x=1}^{K} F_x = 0$
ERROR MODEL	$e \approx N(0,\sigma^2)$ $\sigma^2 = ??$	when $x=0,1$ $E_{xni} = p_{ni1}$ $V_{xni} = p_{ni1}p_{ni0}$
ITEM ESTIMATE	v_i regression of factor score u_n on item i	D_i linear calibration of item i on variable
ITEM ESTIMATE ERROR	$s_i > s_z/N^{1/2}$ $s_z^2 = \sum_{n=1}^{N}\sum_{i=1}^{L}(z_{ni}-u_n v_i)^2/NL$	$s_i = \left[\sum_{n=1}^{N} p_{ni1}p_{ni0}\right]^{-1/2}$
PERSON ESTIMATE	u_n score predicted for person n on factor	B_n linear measure of person n on variable
PERSON ESTIMATE ERROR	$s_n > s_z/L^{1/2}$ $s_z^2 = \sum_{n=1}^{N}\sum_{i=1}^{L}(z_{ni}-u_n v_i)^2/NL$	$s_n = \left[\sum_{i=1}^{L} p_{ni1}p_{ni0}\right]^{-1/2}$
RESIDUAL STATISTICS	$E_e = 0$ $V_e \approx s_z^2 > \sigma^2$	$E\{x_{ni}-p_{ni1}\} = 0$ $V\{x_{ni}-p_{ni}\} = p_{ni1}p_{ni0}$
MISFITS	$(z_{ni}-u_n v_i)^2 >> s_z^2 > \sigma^2$	$(x_{ni}-p_{ni1})^2 >> p_{ni1}p_{ni0}$
TO SEEK NEXT VARIABLE	subtract $u_n v_i$ from all z_{ni} for next data $(z_{ni}-u_n v_i)$ includes residual noise	select only x_{ni} of misfitting items avoids residual noise